

Generating Transition Paths by Langevin Bridges

Henri Orland*

Institut de Physique Théorique, CEA, IPhT

CNRS, URA2306,

F-91191 Gif-sur-Yvette, France

Abstract

We propose a novel stochastic method to generate paths conditioned to start in an initial state and end in a given final state during a certain time t_f . These paths are weighted with a probability given by the overdamped Langevin dynamics. We show that these paths can be exactly generated by a non-local stochastic differential equation. In the limit of short times, we show that this complicated non-solvable equation can be simplified into an approximate stochastic differential equation. For longer times, the paths generated by this approximate equation can be reweighted to generate the correct statistics. In all cases, the paths generated by this equation are statistically independent and provide a representative sample of transition paths. In case the reaction takes place in a solvent (e.g. protein folding in water), the explicit solvent can be treated. The method is illustrated on the one-dimensional quartic oscillator.

*Electronic address: henri.orland@cea.fr

I. INTRODUCTION

The problem of finding the pathway of chemical or biological reactions is of utmost importance for the understanding of their underlying mechanisms, as it allows to have better control on these reactions [1]. For instance, in the realm of proteins, understanding the pathway between the unfolded state and the native state, or between two native states of the protein (allostery) may help prevent certain reactions or on the contrary favor them. Recent progress in single molecule experiments have allowed to monitor the spontaneous thermal folding and unfolding of single proteins, or the force induced unfolding of proteins [2–4].

In the following, we will study the spontaneous or the driven transition between an initial state denoted A and a final state denoted B.

This problem has been addressed mainly by stochastic methods which start from an initial path and deform it by sampling the vicinity of the path. These are the path sampling methods [5–7]. The main drawback of these methods is that they are time consuming, and they generate strongly correlated trajectories. As a consequence, the space of sampled trajectories depends strongly on the initial used path. The same kind of problem exists for the Dominant Pathway method [8, 9], where the minimal action path depends strongly on the initial guess.

From now on, we assume that the system is driven by stochastic dynamics in the form of an overdamped Langevin equation

$$\frac{dx}{dt} = -\frac{1}{\gamma} \frac{\partial U}{\partial x} + \eta(t) \quad (1)$$

For the sake of simplicity, we illustrate the method on a one-dimensional system, the generalization to higher dimensions or larger number of degrees of freedom being straightforward. In this equation, $x(t)$ is the position of a point at time t in a potential $U(x)$, γ is the friction coefficient, related to the diffusion constant D through the relation $D = k_B T / \gamma$, where k_B is the Boltzmann constant and T the temperature of the thermostat. In addition, $\eta(t)$ is a Gaussian white noise with moments given by

$$\langle \eta(t) \rangle = 0 \quad (2)$$

$$\langle \eta(t)\eta(t') \rangle = \frac{2k_B T}{\gamma} \delta(t - t') \quad (3)$$

It is well known that the probability distribution $P(x, t)$ for the particle to be at point x at time t is given by a Fokker-Planck equation [10]

$$\frac{\partial P}{\partial t} = D \frac{\partial}{\partial x} \left(\frac{\partial P}{\partial x} + \beta \frac{\partial U}{\partial x} P \right) \quad (4)$$

where $\beta = 1/k_B T$ is the inverse temperature. In this one dimensional model, the initial state A is characterized by its position x_0 at time 0 and the final state B by its position x_f at time t_f . This equation is thus to be supplemented by a boundary condition $P(x, 0) = \delta(x - x_0)$ where x_0 is the initial position of the particle.

It is convenient to go to the Schrödinger representation, by defining

$$\Psi(x, t) = e^{\beta U(x)/2} P(x, t)$$

The function $\Psi(x, t)$ satisfies the imaginary time Schrödinger equation

$$\frac{\partial \Psi}{\partial t} = \frac{k_B T}{\gamma} \frac{\partial^2 \Psi}{\partial x^2} - \frac{1}{4\gamma k_B T} V(x) \Psi(x) \quad (5)$$

with

$$V(x) = \left(\frac{\partial U}{\partial x} \right)^2 - 2k_B T \frac{\partial^2 U}{\partial x^2} \quad (6)$$

Using the standard notations of quantum mechanics, one can conveniently write

$$P(x_f, t_f | x_0, 0) = e^{-\beta(U(x_f) - U(x_0))/2} \langle x_f | e^{-t_f H} | x_0 \rangle \quad (7)$$

where the Hamiltonian H is given by

$$H = -\frac{k_B T}{\gamma} \frac{\partial^2}{\partial x^2} + \frac{1}{4\gamma k_B T} V(x) \quad (8)$$

In eq.(7), we have denoted by $P(x_f, t_f | x_0, 0)$ the probability for a particle to start at x_0 at time 0 and end at x_f at time t_f , to emphasize the boundary conditions.

It is well-known that the ground state of H , which has 0 energy, is $\Psi_0(x) = e^{-\beta U(x)/2} / \sqrt{Z}$ where Z is the partition function of the system, and all eigenstates Ψ_α of H have strictly positive energies $E_\alpha > 0$. The spectral expansion of P can be written as

$$P(x_f, t_f | x_0, 0) = \frac{e^{-\beta U(x)}}{Z} + \sum_{\alpha \neq 0} e^{-t_f E_\alpha} P_\alpha(x_f, x_0)$$

We see that for large t_f the system converges to the Boltzmann distribution, and that its relaxation time is given by the inverse of the first eigenvalue $\tau_R = 1/E_1$. In systems with high energy barriers, such as proteins, the gap E_1 may be very small, and consequently the time τ_R which in this case is identified with the folding time, can be very long.

Using the Feynman path integral representation, we may thus write eq.(7) as [11]

$$P(x_f, t_f | x_0, 0) = e^{-\beta(U(x_f) - U(x_0))/2} \int_{(x_0, 0)}^{(x_f, t_f)} \mathcal{D}x(t) \exp \left(-\frac{1}{4k_B T} \int_0^{t_f} dt \left(\gamma \dot{x}^2 + \frac{1}{\gamma} V(x) \right) \right) \quad (9)$$

In the following, we will be mostly interested in problems of energy or entropy barrier crossing, which are of utmost importance in many chemical, biochemical or biological reactions. As we already mentioned before, the archetype of such reactions is protein folding, a model we will use in the rest of this paper. A protein is a small biopolymer, which essentially may exist in two states, namely the native state (with biological activity) and the denatured state (with no biological activity) [12]. The protein being a small system (up to a few hundred amino-acids), it never stays in one of the two states, but rather makes rare stochastic transitions between the two states (see fig.1). The picture which emerges is that of the system staying for a long time in one of the two states and then making a rapid transition to the other state.

It follows that for most of the trajectory, the system makes uninteresting stochastic oscillations in the well, and can be described by normal mode analysis. Rarely, there is a very short but interesting physical phenomenon, which is the fast transition from one minimum to the other.

This picture has been confirmed by single molecule experiments [2, 3], where the waiting time in one state can be measured, but the time for crossing from one state to the other is so short that it cannot be resolved. This scenario has also been confirmed recently by very long millisecond molecular dynamics simulations [13] which for the first time show spontaneous thermal folding-unfolding events.

According to Kramers theory, the total transition time τ_K (waiting + crossing) scales like the exponential of the barrier energy

$$\tau_K \sim e^{\Delta E/k_B T}$$

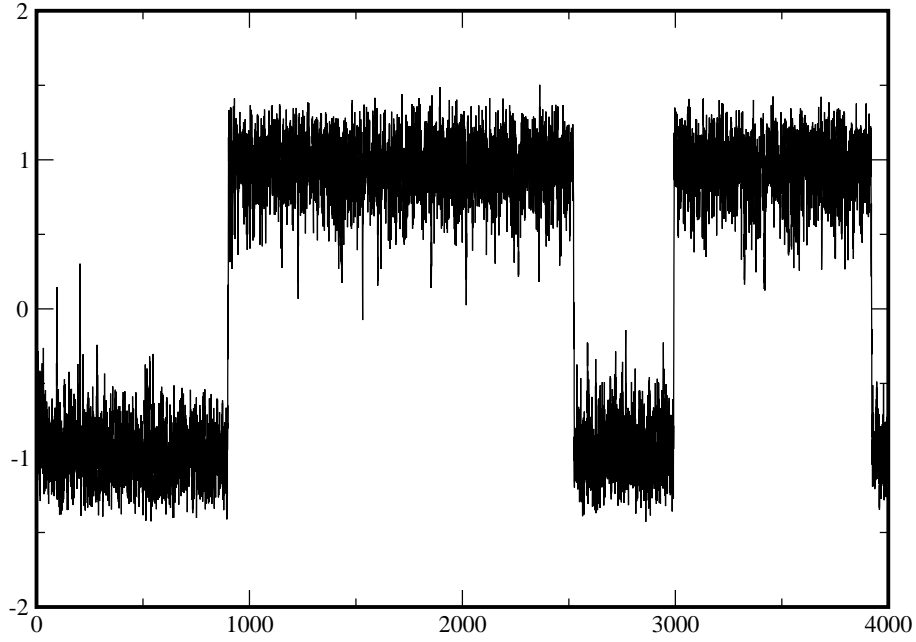


Figure 1: A long Langevin trajectory in the double-well.

The “Kramers time” τ_K is the sum of two times:

- the waiting time in the potential well
- the crossing time over the barrier τ_C

It is well known that the crossing time τ_C is small compared to τ_K and indeed, Hummer [14] and subsequently Szabo [15] have shown that

$$\tau_C \sim \ln \frac{\Delta E}{k_B T} \ll \tau_K$$

These Kramers and crossing times are averages. In fact, these times are distributed (random variables) and single molecule experiments or long molecular dynamics simulations allow to compute their probability distributions.

However, it seems a bit wasteful to simulate proteins over huge time scales (milliseconds), during which only small conformational vibrations occur, just to observe interesting physical

crossing events which occur very rarely, on the sub-microsecond scale.

The goal of this paper is to show how one can generate a representative sample of transition paths, starting in state A at time 0 and ending in state B at some arbitrary time t_f . The typical times of interest are not the (long) folding times, but rather the (very short) transition or barrier crossing times. In mathematical terms, we are looking for the paths starting from A at time 0 and conditioned to end in state B at time $t_f \ll \tau_K$.

II. THE CONDITIONAL PROBABILITY

Using the path integral representation of eq.(9), we see that the probability for a path $\{x(t)\}$ starting at x_0 at time 0, to end at x_f at t_f is given by

$$P(\{x(t)\}) = \frac{1}{A} e^{-\beta(U(x_f)-U(x_0))/2} \exp \left(-\frac{1}{4k_B T} \int_0^{t_f} dt \left(\gamma \dot{x}^2 + \frac{1}{\gamma} V(x) \right) \right) \quad (10)$$

where

$$A = \int dx_f e^{-\beta(U(x_f)-U(x_0))/2} \int_{(x_0,0)}^{(x_f,t_f)} \mathcal{D}x(t) \exp \left(-\frac{1}{4k_B T} \int_0^{t_f} dt \left(\gamma \dot{x}^2 + \frac{1}{\gamma} V(x) \right) \right) \quad (11)$$

The conditional probability over all paths starting at x_0 at time 0 and ending at x_f at time t_f , to find the system at point x at an intermediate time t is given by

$$\mathcal{P}(x, t) = \frac{1}{P(x_f, t_f | x_0, 0)} Q(x, t) P(x, t)$$

where

$$P(x, t) = P(x, t | x_0, 0)$$

$$Q(x, t) = P(x_f, t_f | x, t)$$

The equation satisfied by P is given by (4), whereas that for Q is given by

$$\frac{\partial Q}{\partial t} = -D \frac{\partial^2 Q}{\partial x^2} + D\beta \frac{\partial U}{\partial x} \frac{\partial Q}{\partial x} \quad (12)$$

It follows easily that the equation for the conditional probability $\mathcal{P}(x, t)$ is given by

$$\frac{\partial \mathcal{P}}{\partial t} = D \frac{\partial}{\partial x} \left(\frac{\partial \mathcal{P}}{\partial x} + \frac{\partial}{\partial x} (\beta U - 2 \ln Q) \mathcal{P} \right)$$

Comparing this equation with the initial Fokker-Planck (4) and Langevin (1) equations, one sees that it can be obtained from a Langevin equation with a modified potential

$$\frac{dx}{dt} = -\frac{D}{k_B T} \frac{\partial U}{\partial x} + 2D \frac{\partial \ln Q}{\partial x} + \eta(t) \quad (13)$$

This equation has been previously obtained using the Doob transform [16] and is known in the probability literature as a *Langevin bridge*: the paths $\{x(t)\}$ generated by (13) are conditioned to end at (x_f, t_f) . It is the new term in the Langevin equation that guarantees that the trajectory starting at $(x_0, 0)$ will end at (x_f, t_f) .

Using eq.(7) for Q , one can write equation (13) as

$$\frac{dx}{dt} = 2 \frac{k_B T}{\gamma} \frac{\partial}{\partial x} \ln \langle x_f | e^{-(t_f-t)H} | x \rangle + \eta(t) \quad (14)$$

Using the analogous of the correspondence principle of quantum mechanics [17], i.e. $\frac{\hbar}{i} \frac{\partial}{\partial x} \rightarrow p$, this equation can also be rewritten in the form

$$\frac{dx}{dt} = \langle \dot{x}(t) \rangle + \eta(t) \quad (15)$$

where by definition

$$\langle \dot{x} \rangle = \frac{1}{\langle x_f | e^{-(t_f-t)H} | x \rangle} \int_{(x,t)}^{(x_f,t_f)} \mathcal{D}x(\tau) \dot{x}(t) \exp \left(-\frac{1}{4k_B T} \int_t^{t_f} d\tau \left(\gamma \dot{x}^2 + \frac{1}{\gamma} V(x) \right) \right) \quad (16)$$

Note that for large time t_f , the matrix element in eq.(14) is dominated by the ground state of H , namely $\langle x_f | e^{-(t_f-t)H} | x \rangle \sim e^{-\frac{\beta}{2}(U(x_f)+U(x))}$ and as expected one recovers the standard (unconditioned) Langevin equation.

Since we have a natural splitting of the Hamiltonian H as $H = H_0 + V_1$ with $H_0 = -\frac{k_B T}{\gamma} \frac{\partial^2}{\partial x^2}$ and $V_1 = V/4\gamma k_B T$, it is convenient to rewrite the above equation as

$$\frac{dx}{dt} = 2 \frac{k_B T}{\gamma} \frac{\partial}{\partial x} \ln \langle x_f | e^{-(t_f-t)H_0} | x \rangle + 2 \frac{k_B T}{\gamma} \frac{\partial}{\partial x} \ln \frac{\langle x_f | e^{-(t_f-t)H} | x \rangle}{\langle x_f | e^{-(t_f-t)H_0} | x \rangle} + \eta(t) \quad (17)$$

Note that the first term in the r.h.s. above is singular at $t = t_f$ and is thus responsible for driving the system to (x_f, t_f) whereas the second one is regular. It follows that the first

term is the only term which can drive the system to (x_f, t_f) , and any approximation which keeps the second term finite for $t = t_f$ will not affect this property.

This nice bridge equation cannot be used "as is", since we don't know how to compute the function Q or equivalently the matrix element in the above equation. There are many ways to approximate this function. It is important however, to preserve detailed balance as well as possible, that the approximation retains the symmetry of the matrix element.

III. THE MODIFIED LANGEVIN EQUATION AND REWEIGHTING

The only approximation we found which remains local in time, i.e. which does not give rise to an integro-differential stochastic equation is the symmetric form of the Trotter approximation, commonly used in quantum mechanics [11]. Indeed, for short times t , a very simple and convenient symmetric approximation for Q is given by

$$e^{-Ht} \sim e^{-tV_1/2} e^{-tH_0} e^{-tV_1/2} + O(t^3) \quad (18)$$

which translates into

$$\langle x_f | e^{-Ht} | x \rangle \sim e^{-\frac{\beta\gamma}{4t}(x_f-x)^2 - \frac{\beta t}{8\gamma}(V(x_f)+V(x))}$$

It would be nice to relate the range of validity of this equation to the spectrum of H . Indeed, as was shown before, the spectrum of H corresponds to all the dynamical times of the system (folding times, transition times, etc...). We have not succeeded in finding such a relation except in the solvable case of the harmonic oscillator. In that case, it can easily be shown that the natural expansion parameter is $t\Delta$ where Δ is the constant gap between the energy levels of H . As mentioned before, in the case of protein folding, the folding time which is the inverse of the first gap of the system can be very long, and we might expect the above approximation to be valid for times much smaller than this time. In particular, this approximation would allow to investigate the crossing times, much shorter than the folding time.

Plugging eq.(18) in eq.(26) we obtain the approximate Langevin bridge equation which in arbitrary dimension (or with arbitrary number of degrees of freedom) reads

$$\frac{d\vec{x}}{dt} = \frac{\vec{x}_f - \vec{x}}{t_f - t} - \frac{1}{4\gamma^2}(t_f - t)\nabla V(\vec{x}) + \vec{\eta}(t) \quad (19)$$

where $\vec{\eta}(t)$ is a white noise vector whose components satisfy the relations (2) and (3) and

$$V(\vec{x}) = (\nabla U)^2 - 2k_B T \nabla^2 U \quad (20)$$

The first term in the r.h.s of eq.(19) is the one which drives the particle to reach x_f at time t_f . The potential which governs this bridge equation is not the original $U(x)$ but rather the effective potential $V(x)$. Note also that the force term is proportional to $(t_f - t)$ and thus becomes small as the particle gets close to its target site.

In order to build a representative sample of paths starting at $(x_0, 0)$ and ending at (x_f, t_f) , one must simply solve this equation for many different realizations of the random noise. Only the initial boundary condition is to be imposed, as the singular term in the equation imposes the correct final boundary condition. An important point to note is that all the trajectories generated by eq.(19) are statistically independent. From a numerical point of view, this means that this equation can be fully parallelized, and from a statistical point of view, it implies that all trajectories can be used in the representative sample. This last important point is to be contrasted with most existing methods where the sample are generated by some stochastic (Monte Carlo) methods which generate highly correlated trajectories.

Before presenting examples of application of this method, let us discuss how to correct for the fact that the total time t_f should be small for the approximation to be valid.

Due to this restriction, the statistic of trajectories is not exact for larger times. Indeed, if eq.(19) were exact, each trajectory would be generated with its correct weight, and if one wanted to calculate observables, one would just have to compute simple white averages over all trajectories. However, as the equation is approximate, one needs to resample the ensemble of trajectories, that is, assign them a new weight. As we will show, the resampling weight is easily obtained.

Indeed, if we consider the sample of trajectories generated using eq.(19) between $(x_0, 0)$ and (x_f, t_f) , the weight of each trajectory should be given by eq.(10). However, it is clear from eq.(19) that, using the Ito prescription, the weight with which it was generated is given by

$$\exp \left(-\frac{\gamma}{4k_B T} \int_0^{t_f} dt \left(\frac{d\vec{x}}{dt} - \frac{\vec{x}_f - \vec{x}}{t_f - t} + \frac{t_f - t}{4\gamma^2} \nabla V(\vec{x}) \right)^2 \right) \quad (21)$$

Up to a normalization, the reweighting factor for a trajectory is thus given by

$$\exp \left(-\frac{\gamma}{4k_B T} \int_0^{t_f} dt \left(\left(\frac{d\vec{x}}{dt} + \frac{1}{\gamma} \nabla U \right)^2 - \left(\frac{d\vec{x}}{dt} - \frac{\vec{x}_f - \vec{x}}{t_f - t} + \frac{t_f - t}{4\gamma^2} \nabla V(\vec{x}) \right)^2 \right) \right) \quad (22)$$

This quantity is easily calculated and allows for a correct evaluation of averages over paths.

This reweighting technique can also be used to generate paths statistically exactly sampled according to eq.(13). Indeed, consider eq.(15). The expectation value $\langle \dot{x}(t) \rangle$ can be computed by generating at each time t an ensemble of (approximate) trajectories starting from the current point x at time t and ending at x_f at time t_f by using eq.(19). By reweighting them using the weights of eq.(22), we can reliably compute $\langle \dot{x}(t) \rangle$ and thus solve eq.(15). Note that this procedure which generates correctly weighted trajectories might seem computationally costly. However, since all trajectories are independent, they can efficiently be generated using massive parallelization.

IV. THE NATIVE STATE

Eq.(18) is in fact not quite valid between non normalizable states like $|x\rangle$ and $|x_f\rangle$, in that it is not true to order $O(t^3)$. However it is true between a normalizable state and $|x\rangle$. Assume that the final state of the system is defined by a probability distribution $\phi(x)$. For instance, for the case of a protein, $\phi(x)$ could represent the Boltzmann weight around the native state of the protein. The probability for the system to start at x at time t and end at time t_f in the native state is given by

$$Q(x, t) = \int dy \phi(y) P(y, t_f | x, t) \quad (23)$$

or using (7)

$$Q(x, t) = \int dy \phi(y) e^{-\beta(U(y) - U(x))/2} \langle y | e^{-(t_f - t)H} | x \rangle \quad (24)$$

where ϕ restricts the integration over y to the vicinity of the native state.

With this definition of Q , it is straightforward to see that eq.(12) and (13) are still valid.

Using the approximation (18) we can write

$$Q(x, t) = e^{\beta U(x)/2} e^{-\frac{t}{2} V_1(x)} \int \frac{dy}{A} \phi(y) e^{-\beta U(y)/2} e^{-\frac{t}{2} V_1(y)} e^{-\frac{\beta \gamma}{4} \frac{(y-x)^2}{t_f - t}} \quad (25)$$

where $A = \sqrt{4\pi(t_f - t)/\beta\gamma}$.

As the function ϕ restricts the integration in (25) to the vicinity of the native state, we can approximate the potential $U(x)$ in this region by a quadratic expansion in terms of the normal modes

$$U(x) \simeq \frac{\omega}{2} (x - x_f)^2$$

It follows that V and V_1 are also quadratic and thus the integral (25) can be performed. Although we will consider only one-dimensional cases in the examples, we present the results for the multi-dimensional case.

Denoting by $\omega_{ij} = \frac{\partial^2 U}{\partial x_i \partial x_j} |_{x_i^f}$ the Hessian matrix of normal modes around the native state, the potential U can be written in that region as

$$U(x) = \frac{1}{2} \sum_{i,j} (x_i - x_i^f) \omega_{ij} (x_j - x_j^f)$$

and thus

$$V(x) = \sum_{i,j} (x_i - x_i^f) \Omega_{ij} (x_j - x_j^f) - 2k_B T \text{Tr } \omega_{ij}$$

where the symbol Tr denotes the trace of the normal mode matrix and

$$\Omega_{ij} = \sum_k \omega_{ik} \omega_{kj}$$

The function Q can be easily calculated as

$$Q(x, t) = e^{-\frac{\beta}{8\gamma} (t_f - t) V(x) - \frac{\beta}{4} \sum_{i,j} (x_i^f - x_i) W_{ij} (x_j^f - x_j)}$$

where

$$W_{ij} = \sum_k D_{ik} \left(I + \frac{t_f - t}{\gamma} D \right)_{kj}^{-1}$$

where I is the unit matrix and

$$D_{ij} = \omega_{ij} + \frac{t_f - t}{2\gamma} \Omega_{ij}$$

The bridge equation becomes then

$$\frac{dx_i}{dt} = \frac{1}{\gamma} \sum_j W_{ij}(x_j^f - x_j) - \frac{1}{4\gamma^2} (t_f - t) \nabla_i V(\vec{x}) + \eta_i(t) \quad (26)$$

V. INCLUDING THE SOLVENT

In many cases, in particular protein folding, one wants to include explicitly the solvent molecules, most often water. It is thus desirable to generate trajectories which are conditioned for the protein coordinates, but not for the water molecules.

Denoting by X_i the set of coordinates of the water molecules, the conditional probability, over all paths starting at $\{x_0, X_0\}$ at time 0 and ending at x_f at time t_f (irrespective of the position of the solvent molecules), for the system to be at $\{x, X\}$ at time t is given by

$$\mathcal{P}(x, X, t) = \frac{1}{\int dX_f P(x_f, X_f, t_f | x_0, X_0, 0)} Q(x, X, t) P(x, X, t)$$

where

$$P(x, X, t) = P(x, X, t | x_0, X_0, 0) \quad (27)$$

$$Q(x, X, t) = \int dX_f P(x_f, X_f, t_f | x, X, t) \quad (28)$$

The coordinates X_f of the solvent are integrated over since the trajectories are not conditioned over the solvent molecules.

Using the method described in the previous sections, the exact generalized Langevin equations satisfied by the coordinates are

$$\frac{dx_i}{dt} = -\frac{D_1}{k_B T} \frac{\partial U}{\partial x_i} + 2D_1 \frac{\partial \ln Q}{\partial x_i} + \eta_i^{(1)}(t) \quad (29)$$

$$\frac{dX_i}{dt} = -\frac{D_2}{k_B T} \frac{\partial U}{\partial X_i} + 2D_2 \frac{\partial \ln Q}{\partial X_i} + \eta_i^{(2)}(t) \quad (30)$$

where D_1 and D_2 are resp. the diffusion constants for protein and water molecules and the Gaussian noises $\eta_i^{(1,2)}(t)$ satisfy the relation

$$\langle \eta_i^{(1,2)}(t) \eta_j^{(1,2)}(t') \rangle = 2D_{1,2}\delta_{ij}\delta(t-t') \quad (31)$$

Consider eq.(28). Let us show that the additional force term vanishes. Indeed,

$$\begin{aligned} Q(x, X, t) &= \int dX_f P(x_f, X_f, t_f | x, X, t) \\ &= \int dX_f P(x_f - x, X_f - X, t_f - t) \end{aligned} \quad (32)$$

because of space and time translation invariance. Due to the integration over X_f in eq.(32), we see that $Q(x, X, t)$ does not depend on X , and thus the new drift term in (30) is absent. Therefore the exact equations for the conditional probability in presence of solvent are

$$\frac{dx_i}{dt} = -\frac{D_1}{k_B T} \frac{\partial U}{\partial x_i} + 2D_1 \frac{\partial \ln Q}{\partial x_i} + \eta_i^{(1)}(t) \quad (33)$$

$$\frac{dX_i}{dt} = -\frac{D_2}{k_B T} \frac{\partial U}{\partial X_i} + \eta_i^{(2)}(t) \quad (34)$$

Using the Trotter approximation (18), these equations become (using vector notations)

$$\frac{d\vec{x}}{dt} = \frac{\vec{x}_f - \vec{x}}{t_f - t} - \frac{1}{4\gamma^2}(t_f - t)\nabla_x V(\vec{x}, \vec{X}) + \vec{\eta}^{(1)}(t) \quad (35)$$

$$\frac{d\vec{X}}{dt} = -\frac{D_2}{k_B T} \frac{\partial U(\vec{x}, \vec{X})}{\partial \vec{X}} + \vec{\eta}^{(2)}(t) \quad (36)$$

where the noises are Gaussian, correlated according to eq.(31).

We thus conclude that in presence of the solvent, the protein is evolved through a modified Langevin equation with the effective potential $V(\vec{x}, \vec{X})$, whereas the solvent molecules are evolved according to the standard Langevin equation in presence of the original potential $U(\vec{x}, \vec{X})$.

Extension of this method to the case of the native state (see previous section) is immediate.

VI. EXAMPLE: THE QUARTIC DOUBLE-WELL

We now illustrate the method on the example of barrier crossing in 1d (quartic potential).

$$U(x) = \frac{1}{4}(x^2 - 1)^2$$

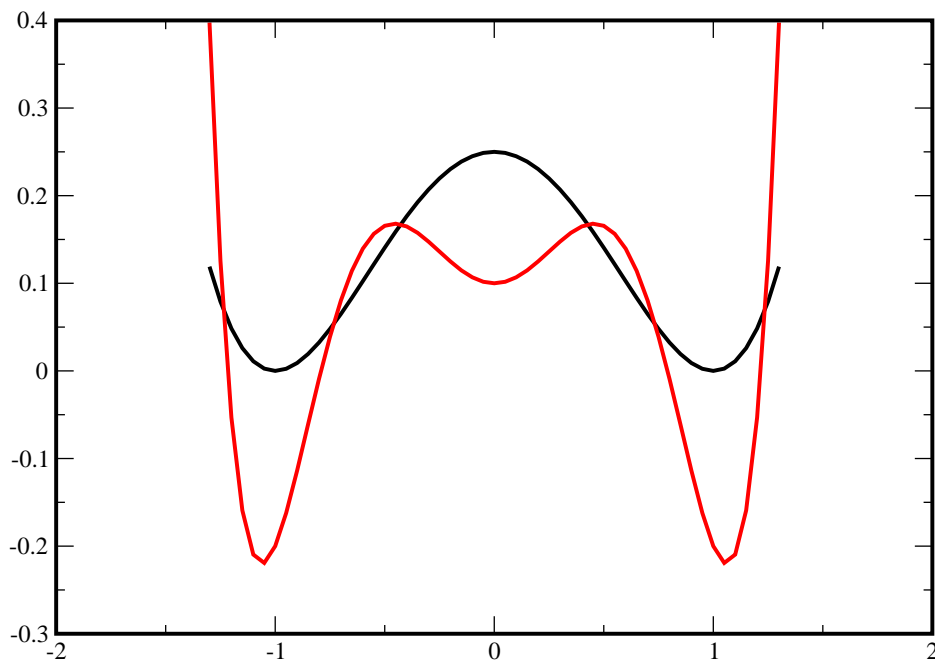


Figure 2: Potential $U(x)$ (in black) and potential $V(x)$ (in red).

This potential has two minima at $x = \pm 1$, separated by a barrier of height $1/4$. Note that at low enough temperature, the potential $V(x)$ has two minima at points close to ± 1 and one minimum at $x = 0$ (from eq.(2)). Note that $V(x)$ is much steeper than $U(x)$ and thus more confining, around its minima.

The model can be solved exactly by solving numerically the Fokker-Planck equation or by diagonalizing the Hamiltonian. All the examples are performed at low temperature $T = 0.05$, where the barrier height is equal to 5 in units of $k_B T$ and the Kramers relaxation time, given by the inverse of the smallest non-zero eigenvalue of H , is equal to $\tau_K = 366.39$.

On fig.3, we present a long trajectory ($t_f = 1000$) obtained by solving the Langevin eq.(1) for a particle starting at $x_0 = -1$ at time 0. The general pattern described in the introduction can be easily checked: the particle stays in the left well for a time of the order of 550, then jumps very rapidly into the right well, where it stays for a time of the order of 200, then jumps back to the left well where it stays again a time equal to about 250.

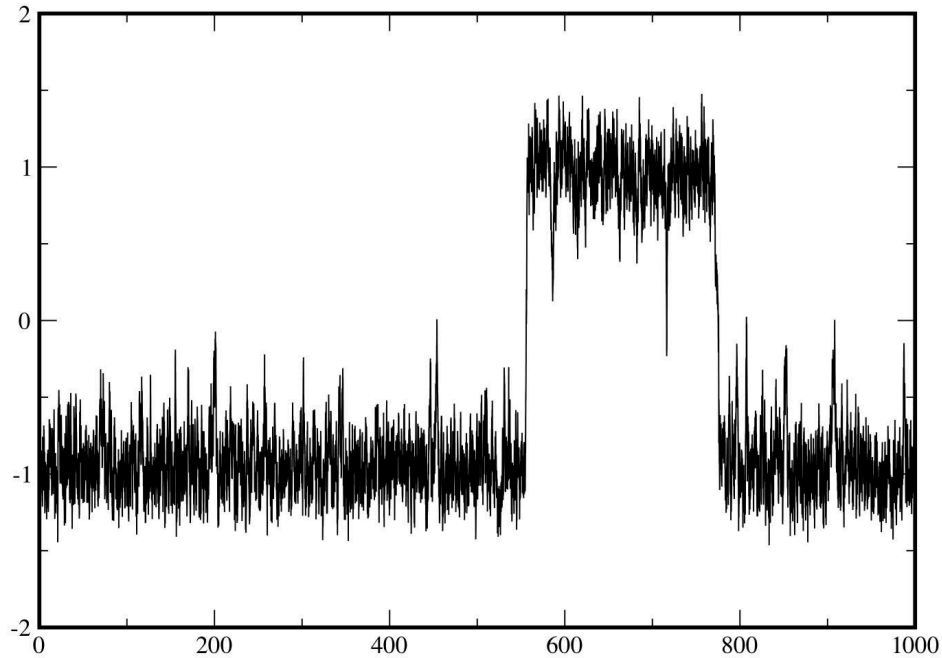


Figure 3: Full Langevin trajectory during time $t_f = 1000$ with 2 transitions between the minima

The two crossings times are very short, and we display an enlargement of the first transition in fig.4.

As can be seen, the crossing time for this specific trajectory is approximately $\tau_C \approx 2.5$, much smaller than the Kramers time.

In fig.5, we plot two examples of two trajectories conditioned to cross the barrier during a time $t_f = 5$. The trajectory in black is obtained by solving the exact bridge eq.(14) by computing exactly (using a spectral decomposition) the matrix element of the evolution operator, while the trajectory in red is obtained by solving the approximate eq.(19) with the exact same sequence of noise $\eta(t)$. In the left figure, the 2 trajectories are barely distinguishable, whereas the agreement is not as spectacular on the right figure.

Next we look at some observables, obtained by averaging over many trajectories.

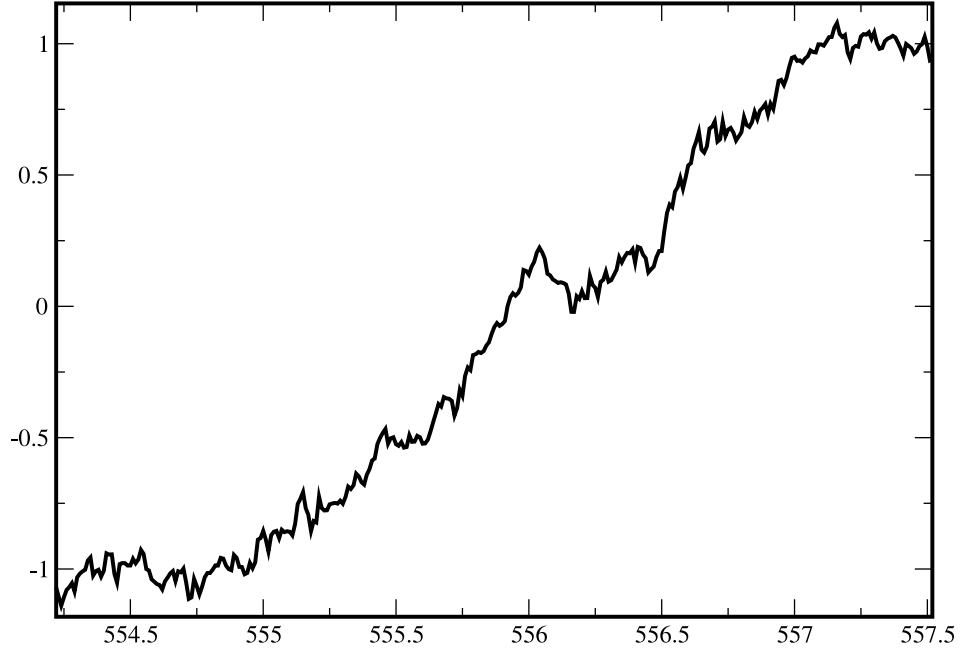


Figure 4: Enlargement of the first transition region

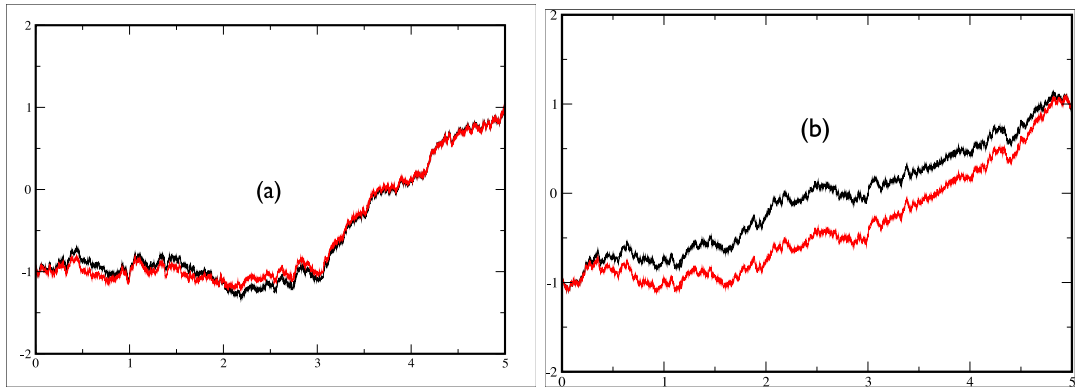


Figure 5: Two sets (a) and (b) of exact trajectories (in black) and approximate trajectories (in red)

In fig.6, we plot: in black the exact average $x(t)$ (obtained by a full expansion over the eigenstates of H), in red the average $x(t)$ over 2000 trajectories obtained by solving eq.(19), and in blue, the average $x(t)$ obtained by reweighting the trajectories according to eq.(22) . Plot (a) is obtained for $t_f = 2$, plot (b) for $t_f = 5$ and plot (c) for $t_f = 10$.

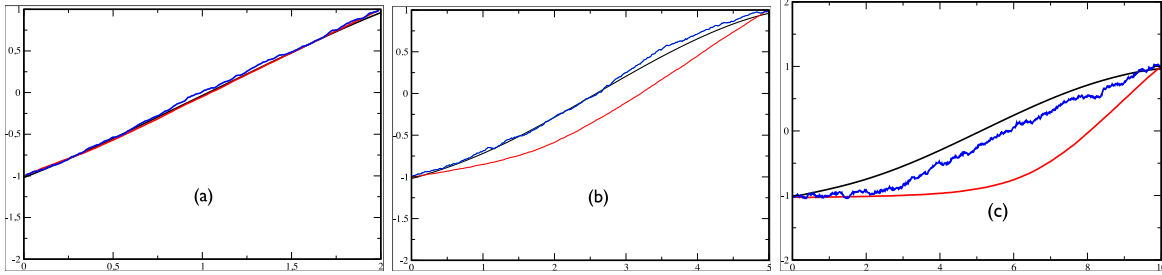


Figure 6: Average position as a function of time for (a) $t_f = 2$, (b) $t_f = 5$, (c) $t_f = 10$. Black curve: exact. Red curve: approximate. Blue curve: reweighted

As expected, we see that the discrepancy between the exact (black) and the approximate (red) average $x(t)$ increases with t_f . For times shorter than the transition time τ_C , the agreement is excellent, whereas for $t_f = 10 > \tau_C$, the agreement is not as good. However, we see that the reweighting procedure, although not perfect, improves drastically the quality of the average for large t_f .

One of the main defects which appears in the approximate theory is the following: In the exact theory, the transition between the 2 minima can take place at any time between 0 and t_f . By contrast, it seems that in the approximate theory, the transition is driven by the final state and takes place only in the end of the trajectory. This effect remains negligible as long as $t_f \lesssim \tau_C$ but becomes important for $t_f > \tau_C$. We illustrate this problem in fig.7 for $t_f = 10$. On the left figure, the exact and approximate trajectories make their transition in the last part of the time, whereas in the right figure, the real trajectory crosses in the beginning while the approximate trajectory still crosses in the last part.

However, as we are interested quantitatively only in the region where the particle crosses the barrier, one can make long runs of approximate trajectories: They will not be good approximations of the real trajectories, except in the end of the trajectory where the transition to the final state occurs.

VII. CONCLUSION

We have presented in this paper a novel method to generate paths following the Langevin overdamped dynamics, starting from an initial configuration and conditioned to end in a given final configuration (point or region of configuration space). We propose an approxima-

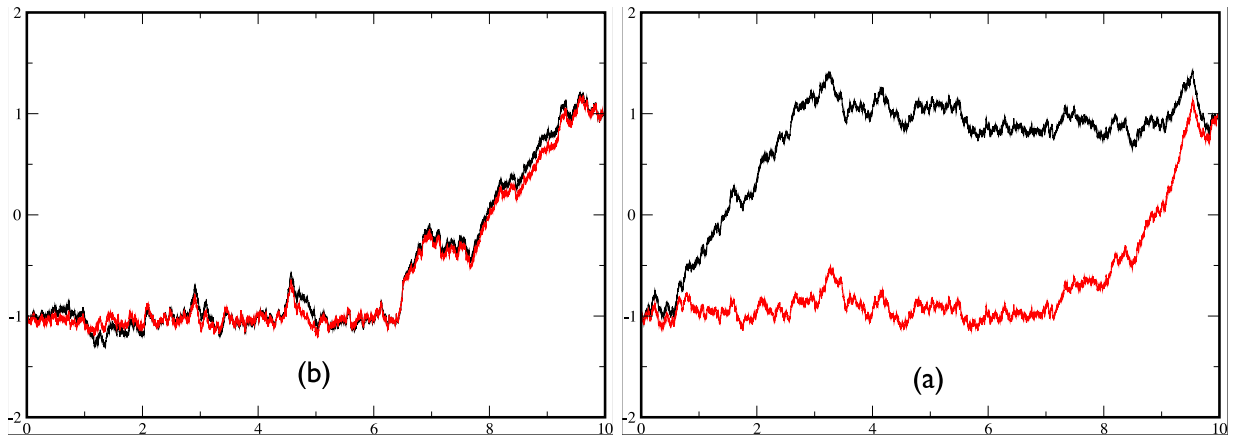


Figure 7: Two sets (a) and (b) of trajectories conditioned to cross the barrier. In black, exact trajectories and in red, approximate trajectories.

tion which is valid for small times. We have not been able to quantify how small should the time be, but the approximate dynamics seems to correctly reproduce the transition through a barrier. The approximate dynamics seems to have a tendency to confine the system in its initial configuration, and to allow for the transition only in the final stages. But this is not really a drawback since if we evolve approximately the system over long times, it will remain close to its initial condition, thus generating unreliable trajectories. However in the latest stages, the system will make its transition to the final state during a short time for which our approximation is reliable.

One of the great advantages of this method is that all generated trajectories are statistically independent. It is thus very easy to generate many of these trajectories using parallel computers. In addition, the trajectories can be reweighted to provide a faithful sample of the exact stochastic dynamics. Finally, this reweighting technique allows for the calculation of the matrix element of the evolution operator, and thus allows for the generation of adequately sampled paths.

The paths generated by our method can also be used either as initial paths to perform Monte Carlo transition path sampling, or as initial conditions for path minimization to determine Dominant Folding Paths.

The method is as simple to implement as ordinary Langevin dynamics, and its application to simple models of protein folding is currently under way.

Acknowledgments

The author wishes to thank M. Bauer and K. Mallick for very useful discussions.

- [1] See e.g. B. Nolting, *Protein Folding Kinetics: Biophysical Methods* (Springer, Berlin) (1999); W.A. Eaton *et al.*, Annual Review of Biophysics and Biomolecular Structure **29** (2000), 327; V. Daggett and A. Fersht, Nature Reviews: Molecular Cell Biology **4** (2003) 497; J.N. Onuchic and P.G. Wolynes, Current Opinion in Structural Biology **14** (2004), 70.
- [2] See e.g. B. Schuler and W.A. Eaton, Current opinion in structural biology **18** (2008) 16-26.
- [3] See e.g. G. Ziv and G. Haran, JACS, **131** (2009) 2942-2947.
- [4] M. Rief, M. Gautel, F. Oesterhelt et al., Science, **276** 5315 (1997) 1109-1112.
- [5] C. Dellago, P. G. Bolhuis, and D. Chandler, J. Chem. Phys. **108** (1998) 9236; P.G. Bolhuis *et al.*, Ann. Rev. Phys. Chem. **53** (2002) 291.
- [6] C. Dellago, P. G. Bolhuis, P. L. Geissler, Adv. Chem. Phys. **123** (2002) 1.
- [7] R. Olender, R. Elber, J. Chem. Phys. **105** (1996) 9299; P. Eastman, N. Gronbech-Jensen and S. Doniach, J. Chem. Phys. **114** (2001) 3823; W.N. E , W.Q. Ren and E. Vanden-Eijnden, Phys. Rev. **B 66** (2002) 052301.
- [8] P. Faccioli, M. Sega, F. Pederiva and H. Orland, Phys. Rev. Lett. **97** (2006) 108101.
- [9] M. Sega, P. Faccioli, F. Pederiva, G. Garberoglio and H. Orland, Phys. Rev. Lett. **99** (2007) 118102.
- [10] N.G. Van Kampen, Stochastic Processes in Physics and Chemistry (North-Holland Personal Edition, 1992); R. Zwanzig, Nonequilibrium Statistical Mechanics (Oxford University Press, 2001).
- [11] R.P. Feynman and A.R. Hibbs, Quantum Mechanics and Path Integrals (McGraw-Hill,1965).
- [12] T.E. Creighton, Proteins: Structures and Molecular Properties (W. H. Freeman, 1992).
- [13] D.E. Shaw,P. Maragakis, K. Lindorff-Larsen et al., Science, **330** (2010) 341-346.
- [14] G. Hummer, J. Chem. Phys. **120** 2 (2003) 516-523.
- [15] A. Szabo, private communication.
- [16] J. L. Doob, Bull. Soc. Math. France **85** (1957), 431-458; P. Fitzsimmons, J. Pitman and M. Yor, Seminar on Stochastic Processes: " Markovian bridges: construction, Palm interpretation,

and splicing" (1992) 101-134.

[17] C. Cohen-Tannoudji, B. Diu and F. Laloe, Quantum mechanics (Wiley-Interscience, 2006).